A bi-directional task-based corpus of learners' conversational speech

A bi-directional task-based corpus of learners' conversational speech

Maria Luisa Garcia Lecumberri[1]
Martin Cooke[2,1]
Mirjam Wester[3]

[1] University of the Basque Country, Vitoria, Spain
[2] Ikerbasque (Basque Science Foundation), Bilbao, Spain
[3] University of Edinburgh, Edinburgh, UK

Abstract

This paper describes a corpus of task-based conversational speech produced by English and Spanish native talkers speaking English and Spanish as both a first and a second language. For cross-language comparability, speech material was elicited using a picture-based task common to each native language group. The bi-directionality of the corpus, stemming from the use of the same speakers and the same language pairing, makes it possible to separate native language factors from the influence of speaking in a first or second language. The potential for studying first language influences and non-native speech using the corpus is illustrated by means of a series of explorations of acoustic, segmental, suprasegmental and conversational phenomena. These analyses demonstrate the breadth of factors that are amenable to investigation in a conversational corpus, and reveal different types of interactions between the first language, the second language and non-nativeness.

Keywords: L2 speech, DIAPIX corpus, conversational speech

1. Introduction

Many language learners would attest that speaking in a non-native language can be an arduous task, requiring the simultaneous juggling of a series of goals, from correct segmental realisations of second language (L2) sounds to appropriate patterning of intonation and stress. Understanding this potent cocktail of challenges has both theoretical and pedagogical implications, and has been the subject of much study.

Much work on non-native speech characteristics has focused on segments and how they are influenced by the sounds of a speaker's native language (L1) (Best, 1995; Flege, 1995). Attempts to find general features of non-native speech have also looked at suprasegmentals and fluency measures, including stress (Kormos & Dénes, 2004), pausing (Riazantseva, 2001), fundamental frequency (Kang et al., 2010), speech rate (Guion et al., 2000; Kormos & Dénes, 2004; Munro & Derwing, 1998), and voice quality (Munro et al., 2010).

One approach to examining what characterises non-nativeness in a learner's speech patterns is to focus on differences between their L1 and L2 productions (e.g., Derwing et al., 2009; Flege, 1987; Flege & Eefting, 1987; Riazantseva, 2001; Rose, 2013). However, in such studies it is difficult to distinguish the features that characterize the act of speaking as a non-native from those that originate in the differences between the first and second languages themselves. For instance, due to differences in syllabic and lexical structure between languages, it may be problematic to define a speech rate measure that can be compared across languages; consequently, it is unclear to what degree any changes in speech rate observed in non-native speech are due to non-nativeness or simply due to language differences. The solution we propose and elaborate in this paper is the use of *bidirectional* corpora involving a language pair spoken by two cohorts of speakers. For one group, the first language of the pair is their L1 and the second their L2; for the other group, these roles are reversed. Specifically, the current study looks at native English speakers who are learners of Spanish producing speech in both their L1 (English) and their L2 (Spanish), and compares their productions to native Spanish speakers who are learners of English producing speech in their L1 (Spanish) and L2 (English). By including both L1 and non-native productions in both languages from the same speaker cohorts, we hope to identify general factors of speaking non-natively without a confounding influence from the language pairing.

To strengthen any comparison of native and non-native speech it is desirable to elicit speech material of a similar complexity in the two languages. While read (i.e. scripted) speech has the advantage that the material can be controlled much more precisely than in conversational speech, there are several reasons to prefer the latter. Spontaneous conversational speech is more representative of the everyday communicative situation in which both native and non-native talkers find themselves, and differs in many ways from read speech e.g., in the choice of words, syntactic structures, frequency and type of hesitations, speech rate, pause types, pause structure, intonation, and acoustic characteristics (Blaauw, 1994; Ernestus et al., 2015; Howell & Kadi-Hanifi, 1991; Laan, 1997; Nakamura et al., 2008). Consequently, analyses using spontaneous speech provide more realistic answers to questions concerning non-native speaking. Another reason to avoid read speech is to sidestep the introduction of orthographic influences during the process of eliciting native and non-native speech.

Spontaneous speech of a similar complexity in two languages can be elicited using a picture description task. We chose an existing paradigm, Diapix, developed for American

English by Van Engen et al. (2010). Diapix requires pairs of participants to spot the differences in simplified pictures such as those illustrated in Figure 1. The Diapix task generates significant amounts of spontaneous, topic-driven speech as participants describe to the other participant what they see in their specific version of the picture. While lacking the control of a read speech task, items represented in the Diapix picture collection can be chosen in such a way as to increase the likelihood of speakers producing words of interest e.g. minimal pairs. For the current corpus we adapted the UK version of Diapix (DiapixUK; Baker & Hazan, 2011). DiapixUK was shown in Baker & Hazan (2011) to lead to similar amounts of speech across both members of the pair. Other claimed advantages include the absence of a learning effect of completing more than one picture in a session, and a similar level of difficulty across picture pairs.

Section 2 describes the collection of the bidirectional corpus of L1/L2 speech for English and Spanish, which we refer to as the “DiapixFL corpus”. The range of research possible using the DiapixFL corpus is illustrated by a number of investigations into non-native speech that examine conversational phenomena such as pausing, elongations and incomplete words (section 3.1), speech rate (section 3.2), acoustic parameters such as energy and spectral tilt (section 3.3), prosodic factors involving fundamental frequency and its range (section 3.4), as well as a segmental measure based on corner vowels (section 3.5). Each analysis attempts to measure the separate influence of two factors: the L1 of the speaker, and whether they are talking natively or non-natively.

An initial report on the DiapixFL corpus was presented in Wester et al. (2014).

2. The DiapixFL corpus

2.1 Materials

The full set of DiapixUK materials (Baker & Hazan, 2011) consists of 12 picture pairs, each belonging to one of three themes: Beach, Farm or Street. For our materials, we selected two picture pairs per theme i.e. six in all, to enable each of the three themes to be used in each of the two languages. The two versions of the picture in each pair differ in 12 places. Participants work in pairs and each receives one version of the picture, their task being to find the locations where the pictures differ by describing verbally what they see in their version. Participants are not able to see their interlocutor's picture. The left and middle panels of Figure 1 show a fragment of the Street scene from the original DiapixUK materials as seen by each member of a participant pair. To elicit speech in Spanish, all text was replaced by its Spanish equivalent, as shown in the right hand panel of Figure 1. Other than textual differences, the pictures used to elicit English and Spanish speech were identical.

<figure 1 here>

2.2 Participants

Participants were native English and native Spanish/Basque speakers who were recruited at, respectively, the University of Edinburgh in the UK and the University of the Basque Country in Spain. The participants were all in their second year of university studying either Spanish in Edinburgh or English in the University of the Basque Country. All possessed a CEFR level of B2/C1 for their foreign language (Council of Europe, 2001), ensuring that the level of proficiency across the speakers for these languages was comparable.

At each site six pairs were recorded, i.e., twelve speakers. In each case there were 10 female and 2 male talkers. Speakers were remunerated for their time and effort. Table 1 provides further information about participants' backgrounds. Note that participant S11, though born in Singapore, was judged to have a British English accent. Participants recorded at the University of the Basque Country had a northern peninsular Spanish accent.

Speaker	Gender	L1	Age	Place of birth
S1	F	En	18	Central Scotland
S2	F	En	20	NW England
S3	F	En	19	S England
S4	F	En	19	SE England
S5	F	En	19	NE Scotland
S6	F	En	19	NE Scotland
S7	F	En	20	SE England
S8	M	En	20	SE England
S9	F	En	19	SE England
S10	F	En	19	S England
S11	F	En	20	Singapore
S12	M	En	20	Central Scotland
S13	F	Sp	19	Basque Country
S14	F	Sp	19	Basque Country
S15	M	Sp	23	Burgos (N Spain)
S16	M	Sp	20	Navarra (NE Spain)
S17	F	Sp	19	Basque Country
S18	F	Sp	19	Basque Country
S19	F	Sp	20	Basque Country
S20	F	Sp	21	Basque Country
S21	F	Sp	19	Basque Country
S22	F	Sp	19	Basque Country
S23	F	Sp	19	Basque Country
S24	F	Sp	19	Basque Country

Table 1: Participant details.

2.3 Recording setup and procedure

During a single session a single pair of participants was recorded. Participants were seated at a table in a recording studio with a divider between them. The divider made it impossible for them to see each other's picture, but they could see each other. Recordings were made with both close-talking and table microphones. Signals were digitised at 22.05 kHz with 16-bits amplitude quantisation. At the University of Edinburgh, the close-talking microphone was a DPA4066 omnidirectional headset microphone, while the table microphone was a Sennheiser model MKH800 P48. At the University of the Basque Country the models were AKG 4500 and Sennheiser Me-3 respectively.

Each pair was asked to spot the 12 differences in six pictures. Three of the pictures - one each of the Beach, Farm or Street themes - were in Spanish, the other three in English. The pictures for each theme were different in Spanish and English. The order of the language spoken was alternated, with half the pairs starting in Spanish, the other half in English. After completing three pictures the language was switched for the final three pictures. In this way the recording context (e.g., speaker locations and microphone positioning) was the same when speaking both languages. Prior to starting the recording, the DiapixUK training picture pair of the Park scene was used to familiarise the participants with the task.

2.4 Annotation

For each of the two languages in the recordings, a native speaker of that language annotated all speech material (e.g., the English transcriber annotated the English speakers speaking natively and the Spanish speakers speaking non-natively). Subsequently, a balanced bilingual speaker who was also an expert in English and Spanish phonetics cross-checked the two sets of annotations for correctness and consistency. Annotation involved identification and labelling of turn construction units (TCUs; section 3.1) and orthographic transcription of those TCUs containing speech. The multilevel annotation tool Mtrans (Villegas et al., 2011) was used for TCU/orthographic annotation. Corner vowels in a subset of the corpus were transcribed using Praat (Boersma & Weenink, 2016) by different annotators (see section 3.5).

3. Illustrative findings

This section presents more detailed information on the composition of the DiapixFL corpus and summarises findings with respect to the role of first language and the differences between native and non-native speaking. A range of acoustic, segmental, suprasegmental and turn-type parameters for each individual speaker when conversing natively or non-natively is examined.

Due to the gender imbalance amongst the participants, and to avoid gender-based normalisations of parameters such as F0, the analyses presented here used only material from the female talkers only, i.e., 10 English and 10 Spanish talkers.

In Figure 2 and subsequent figures, speakers are numbered from 1-10 in each of the two languages, and values when speaking natively and non-natively are distinguished. Note that the abbreviations En and Sp signify the native language of the participants as opposed to the language being spoken, while the terms N and NN denote whether the language is being spoken natively or non-natively. For example, En NN identifies the condition where the English native speakers were speaking Spanish.

Statistical analyses are based on mixed-effects ANOVAs with a within-subjects factor of nativeness (N versus NN speech) and a between-subjects factor of L1 (En versus Sp). Comparisons of levels with and across factors make use of Fisher's Least Significant Differences (LSD). Table 2 provides a statistical summary of main effects and interactions for the parameters described below.

Parameter	L1	Nativeness	L1 x Nativeness	Figure
speech proportion	91.5 (0.74)***	56.8 (0.57)***	9.8 (0.19)***	2
pause proportion	14.3 (0.41)**	72.1 (0.35)***	7.1 (0.05)*	2
filled pause proportion	8.0 (0.23)*	9.7 (0.15)**	p=0.07	2
nonspeech proportion	23.1 (0.47)***	p=0.34	p=0.89	2
elongations	6.3 (0.21)*	40.8 (0.34)***	p=0.06	3
incomplete words	4.5 (0.18)*	p=0.12	p=0.76	3
normalised speech rate	6.0 (0.18)*	167 (0.76)***	p=0.81	4
energy	11.4 (0.34)**	p=0.12	p=0.08	5
spectral tilt	p=0.47	p=0.35	p=0.11	5
voicing	54.6 (0.63)***	p=0.79	19.6 (0.32) ***	5
F0 mean	p=0.47	p=0.54	p=0.39	6
F0 deviation	p=0.07	8.8 (0.05)**	p=0.29	6
min F0	p=0.62	4.6 (0.02)*	p=0.65	6
max F0	p=0.20	p=0.40	p=0.45	6

Table 2: Statistical summary of main effects for the factors L1 and nativeness and their interaction. For significant effects, columns list the F(1,18) value with η2 in parentheses, alongside an indication of significance level (***: $p < .001$; **: $p < .01$; *: $p < .05$); for non-significant effects, actual p values are shown.

3.1 Turn types

For the analysis of speaker turns, the entire contribution of each speaker when carrying out the task was used. In addition to the words themselves, elements corresponding to pausing, external noise, and extralinguistic features such as in-breaths were annotated using the symbols shown in Table 3. Segments of speech, non-speech vocalisations and other events are referred to as turn types. In the annotation of pausing, we distinguish between three types: (i) filled pauses (an entire pause that includes a filler such as “uh”, “um” or “er”); (ii) unfilled pauses which are pauses during a speaker's turn; and (iii) silence on the part of the listener when the interlocutor is talking.

Symbol	Meaning
-	silence (listening to interlocutor)
#	non-speech vocalisation (e.g., laughter)
@	external noise
*	incomplete word
:	elongation
%	filled pause
+	unfilled pause
<	in-breath
$	code-switching
?	transcriber unsure of utterance

Table 3: Transcription symbols and descriptions.

In absolute terms, TCUs identified as speech captured across all participants and conditions totalled just over 5.5 hours, representing 6.88 minutes on average per talker/ condition. The least voluble talker produced 3.50 minutes of speech material, while the most voluble generated 14.73 minutes. Although more speech was generated when talking non-natively (7.77 min compared to 5.99 min when speaking natively), the length of session (corresponding to the time required to complete the task) was longer when talking non-natively (27.10 min vs 17.30 min talking natively), resulting in a lower proportion of speech turns in non-native speech.

Figure 2 plots individual speaker and cohort means for the percentages of speech and non-speech turns, and turns consisting of pauses and filled pauses; together, these turn types make up over 94% of all turns. The proportion of the session containing speech turns is higher in native speech (for statistical details see table 2), and as a cohort Spanish speakers produced more speech than English speakers. The proportion of both types of pause is higher in non-native speech. This is a typical correlate of hesitant speech that has been observed previously for non-native speech (Riazantseva, 2001). Again, L1

differences were observed, with Spanish talkers producing proportionally fewer of the two pause types. Mild but significant interactions between L1 and nativeness were observed for speech and pauses, with a similar tendency for filled pauses; in these cases the difference between the native and non-native conditions was always larger for the Spanish cohort.

<figure 2 here>

There were no clear intra-speaker correlations between the amount of speech produced in the native and in the non-native language. For instance, English native speaker number 5 produced the largest amount of speech in her native language but was the third least productive in Spanish. This may be related to non-native language competence level, highlighting a possible pedagogical application of the present corpus.

As part of the annotation of turn types, words which were abnormally elongated or incomplete (e.g., cut off during articulation due to hesitation as opposed to showing elisions typical of weakening processes) were marked. Figure 3 plots the number of elongated and incomplete words produced per minute of the corpus for each speaker. Spanish speakers produced a greater number of elongated words than English speakers when speaking natively (15 versus 6 per minute). Both speaker groups significantly increased the quantity of elongations in non-native speech. As for pausing, elongations are characteristic of hesitations; in the case of elongations when speaking non-natively, hesitations might also reflect the higher cognitive burden of retrieving L2 words. The increase in elongations when speaking non-natively was significantly greater for the English group (170% versus 30%), possibly due to the effect of attempting to reproduce what is perceived as normative in Spanish, combined with the non-nativeness factor.

Interestingly, the number of incomplete words was not affected by nativeness, suggesting that lack of completion is influenced mainly by turn-taking phenomena such as interruptions rather than as a consequence of speaking in a non-native language, and that talkers retrieve complete lexical candidates. For reasons that are currently unclear, English speakers in this corpus produced significantly more incomplete words than Spanish speakers.

3.2 Speech rate

For the analysis of speech rate, all turn construction units marked as speech were used. Raw speech rate is an unreliable measure for across-language comparison since the differing structures of each language (e.g., frequencies of monosyllabic words) will influence the number of words a typical speaker will produce in a given time. The solution we adopted was to normalise the speech rate so that the average rate is identical for both languages when spoken natively. In DiapixFL English spoken natively results in 231 words-per-minute (WPM), while for Spanish the comparable figure is 219 WPM, so the normalised WPM measure results from multiplying the WPM for English (both spoken natively and non-natively) by 219/231. The resulting normalised WPM measure is shown in Figure 4.

<figure 4 here>

This figure demonstrates clearly that non-native speech is produced at a significant slower rate (table 2), with a larger reduction for the English group (144 versus 188 WPM). Slower speech is a manifestation of tentative speech, and in particular of non-native speech, where it has been related to linguistic competence (Rose, 2013). These results also reflect an interaction between non-nativeness and intrinsic language characteristics: native English speech rate in the L2 decreases as a function of non-nativeness but also due to the greater frequency of polysyllabic words in Spanish; conversely, although native Spanish speakers also display a reduction in speech rate in their L2, this fall is tempered by the presence of more monosyllables in English, their L2.

3.3 Energy, spectral tilt and voicing

Since speakers produced different total amounts of speech, a uniform length subset of material was chosen for each speaker to be used in the acoustic analyses. Using only those turns marked as speech, contributions longer than 1.4 s were selected in order to to avoid very short utterances (e.g., back-channels). Under this constraint, each talker produced at least 82 s of speech while speaking natively and non-natively. Here, a fixed overall duration of 60 s was used for each talker. To avoid bias in the selection of material from any specific stage in the conversation, speech segments were chosen at random. The fragments extracted for energy, spectral tilt and voicing were also employed in the F0 and vowel studies described in sections 3.4 and 3.5 below.

Voicing was extracted in 10 ms frames using Praat, based on those frames where Praat reported an F0 value. Subsequently, the percentage of voiced frames was computed. Log energy across the 1 minute sample was also computed via Praat. However, since the energy values reported here are uncalibrated, we focus solely on the difference in energy when speaking natively and non-natively. Spectral tilt was estimated using custom Matlab code via a linear fit to energies in third-octave bands.

Figure 5 plots these acoustic parameters. As noted above, the difference in overall dB energy between English and Spanish reflects different recording conditions in the two studios. On average, English speakers did not alter their speech level when talking natively and non-natively. However, there is much variation across individual talkers, with four showing a clear reduction in level, and four exhibiting a clear increase. For Spanish, the picture is more consistent: apart from speaker 2, all speakers spoke more quietly when speaking English. Overall, Spanish non-native speech is 2.4 dB lower than Spanish spoken natively [LSD = 2.0]. We interpret this as a characteristic of tentativeness or lack of confidence.

<figure 5 here>

Spectral tilt was unaffected by the language spoken or nativeness. Changes in spectral tilt generally accompany speech spoken under stress or in the presence of noise (see e.g., Cooke et al., 2014, for a review); the lack of changes in tilt when speaking non-natively may suggest that participants were engaged in the task rather than being aware of talking in an L2.

On the other hand, the ratio of voiced to unvoiced frames shows a clear influence of the L1 (table 2), with Spanish displaying a characteristically higher proportion of voiced speech due to its tendency towards a CV syllable structure. Not surprisngly, the effect of speaking non-natively interacts with the L1: non-natively spoken speech adopts the language

characteristics of the language being spoken. Interestingly, neither cohort achieved the voicing ratio of native speech: English talkers produced Spanish with 64% voiced frames as compared to the 77% when spoken natively, while Spanish talkers produced English with 70% voiced frames versus the 57% typical of native English speech.

3.4 Fundamental frequency

Fundamental frequency (F0) estimates were generated using a procedure designed to detect and remove pitch-halving errors, a typical problem in automatic pitch tracking. For each speaker, F0 estimates in each 10 ms frame were provided via Praat, with upper and lower bounds set to 50 and 400 Hz respectively, using all speech from that speaker. These values were then assembled into a histogram, and a 2-mixture Gaussian model was fitted using the expectation-maximisation algorithm. If the resulting pair of frequency estimates (one from each mixture component) differed by more than 50 Hz, the higher value was chosen as the F0 estimate, and any F0 values lower than the frequency at the mid-point of the two frequency estimates were discarded. If the two values were closer than 50 Hz, a single-component Gaussian was fitted to the histogram of F0 values. From the non-discarded F0 values, robust estimates of mean, standard deviation, minimum and maximum F0 were computed after removing outliers, defined as values more than 1.5 times the inter-quartile range below or above the first and third quartile boundaries.

Figure 6 plots the mean, standard deviation, minimum and maximum F0. While no effect of L1 or nativeness is observed for mean F0 (table 2), non-native speech has a slightly smaller range, as measured by the standard deviation of F0 values. No differences are seen in the maximum F0 reached, but natively-spoken speech has a somewhat lower minimum. Although the effect is modest, both the reduction in F0 range and the lower minimum F0 convey an impression of less confident speech which is often found in non-native speaking as well as polite and tentative speech (Bolinger, 1989; Vaissière, 2008).

3.5 Corner vowels

An acoustic analysis of the three corner vowels in each language was carried out with the aim of exploring L1 influences on the production of vowels which have a similar counterpart in the L2 and whose differences may thus not be noticed by learners (Best, 1995; Flege, 1995; Kuhl, 1993). Additionally, we were interested to see to what extent a current sound change in one of the languages is reflected in the other. A fuller report of this study is contained in Wester et al. (2015).

Vowel midpoints were labelled by two native speakers of the respective languages. Labellers were instructed to (i) mark midpoints of the three vowels: /a/, /i/, and /u/; (ii) find at least 3 examples of each vowel per speaker/language; (iii) aim for stressed syllables; (iv) locate clear examples of the vowel (e.g., not conversationally reduced); and (v) avoid following /r/. Frequencies of the first three formants at vowel midpoints were estimated using Praat. A proficient bilingual speaker with a background in acoustic-phonetics checked both the location of all vowel midpoints and the reliability of the formant estimates for speech material in both languages. Some 674 vowel instances were marked, corresponding to a mean of 5.62 tokens per vowel per talker per language condition. All 674 vowel instances were used in the vowel space analysis.

First and second formant locations for the three corner vowels produced by each speaker are shown in Figure 7. When native speakers of each language are compared, they differ mainly in three respects. For /i/, English speakers show more tightly clustered and fronted (higher F2) realisations than Spanish native values. On the other hand, for /a/, English native speakers show much more height dispersion (F1) than seen in Spanish native /a/ productions, which is understandable given that this vowel differs in height across English accents (e.g., it is lower in Northern British English than in Southern English accents (c.f. Wells, 1982). The most noticeable difference between native speech in English and Spanish in our corpus is seen in the front-back dimension for /u/, where native English speakers display highly-fronted (high F2) values, whereas Spanish native /u/ is consistently produced as a back vowel. This fronted realisation of English /u/ is an ongoing sound change that has been amply documented (Harrington et al., 2008; Hawkins & Midgley, 2005; Scobbie et al., 2012) and which differs from the traditional back pronunciation described for Standard British English during the twentieth century (e.g., Gimson, 1964; Wells, 1962).

<Figure 7 here>

The non-native productions of our corpus manifested changes in the three above characteristics. Spanish speakers with more retracted /i/ values in their native language moved slightly to the front of the vowel space when speaking English (speakers 5 & 10); Spanish learners also showed more dispersion in /a/ height when speaking non-natively and, notably, some speakers produced quite fronted /u/ realisations in English (particularly speakers 2 & 4). All of these are clear adjustments towards the current L2 realisations and away from the L1 values.

English speakers of Spanish showed a strong adaptation of /u/ towards Spanish norms, with significant retraction of its articulation (F2 lowering). There are also individual changes in /a/ height which are probably adjustments to what learners perceive is the target quality in Spanish, particularly by speakers with extreme height values; for instance speaker 8 considerably raised this vowel and speaker 10 lowered it by a similar degree.

Our results suggest that adaptation to the values of the target language is very much speaker- and vowel-dependent. Some speakers showed great adjustments for one vowel but not for another whereas other learners maintained their L1 values, which is understandable since canonically these three vowels may be considered to be quite similar in the two languages and similar sounds are usually more impervious to L2 acquisition (e.g., Flege, 1995). It is only in /u/ backing by English learners of Spanish where we find a consistent effect across learners. We speculate that this adaptation is easier both because the difference between realisations of this vowel in the two languages is larger then for the other corner vowels, and because the Spanish vowel exhibits values which exist as regional variants in English, and which therefore the English speakers may be accustomed to noticing and even imitating.

4. Discussion

The aim of the bidirectional DiapixFL corpus is to permit the separation of L1-specific factors from the influence of non-nativeness in the analysis of spoken language learner corpora. There are relatively few previous studies in which talkers' L1 speech production is considered alongside their non-native speech production (Derwing et al., 2009; Rose, 2013) and further compared to native speakers' of the target language (Riazantseva,

2001). These studies lack the final component which makes our corpus symmetrical, namely, both groups of speakers communicating in the others' language. This inter- and intra-group comparison is essential to be able to separate L1-specific and speaker-specific traits from non-native style characteristics. The bilingual corpus for French and German language learners (Trouvain et al., 2016) is an example of a dataset that is similar to DiapixFL in terms of bi-directionality, but differs in that it is mainly read speech and therefore not as representative of natural, conversational language phenomena. Indeed, another key feature of DiapixFL is the fact that it consists of spontaneous L2 speech which, in spite of being conversational in nature, contains a number of comparable structures and lexical items in the two languages. Granlund et al. (2012) provides another example of a corpus that has used the Diapix elicitation technique in two languages, English and Finnish, albeit uni-directionally (Finnish is not present as an L2).

The spectrum of findings presented above illustrates the potential value of the bidirectional approach which adopts the same task and participants speaking in both their L1 and their L2. The foregoing analyses reveal two types of non-native influence. In one, non-nativeness as a factor impacts in a similar way on both cohorts differing in L1. For example, non-nativeness results in a higher proportion of filled and non-filled pauses, more elongations and, after normalising away L1-based differences, a slower speech rate. Further, non-native speech has a reduced F0 range. A second type of influence of speaking non-natively is apparent in L1-by-nativeness crossover interactions, manifest by a shift in some parameter value from native to non-native norms. Naturally, the shifts are seen in opposing directions. One example is the change in the ratio of voiced frames. While the direction of the shift results from the L1-influenced target, the fact that in non-native speech the shift is typically incomplete (that is, falls short of reaching native norms) is a clear marker of non-native competence.

Not surprisingly, individual differences are also seen throughout our analyses of the DiapixFL corpus. It is beyond the scope of the current work to examine these in detail. However, we note that one domain where speakers are most clearly distinguished is in the production of the English corner vowels in the non-native language. Here, two Spanish talkers showed much more fronted (and native-like) /u/ vowels than the rest of the Spanish cohort (Figure 7). It is tempting to interpret this finding as representing differences in L2 competence: these specific speakers also showed some of the smallest reductions in speech rate (Figure 4). However, other potential indicators of advanced L2 acquisition, such as maintenance of the proportion of speech-based turn types as opposed to, say, within-turn pauses, are more ambivalent.

Our analyses to date have focused on a small set of acoustic parameters and a single segment type; however, the DiapixFL corpus has the potential to support studies of a wider set of phenomena e.g., in intonation, rhythm, stress, as well as for a broader set of phonetic features.

5. Conclusions

A symmetric corpus of learner speech material has been presented, along with a range of illustrative acoustic, segmental, suprasegmental and conversational analyses of the speech material in the corpus. The corpus contains task-based conversational speech spoken by the same talkers in both their L1 and their common L2, and is balanced by speech from a similar cohort in which the roles of the L1 and L2 are swapped. The corpus supports studies in which non-native speech features can be determined and separated

from L1-specific factors. The DiapixFL corpus is freely available at http://datashare.is.ed.ac.uk/handle/10283/346.

Acknowledgements: Collection of the DiapixFL corpus was undertaken as part of the EU-funded Future and Emerging Technology project *The Listening Talker* (grant number 256230). Subsequent annotation was supported by the Spanish Ministry MICIN DIACEX project and the Basque Government *Consolidado* grant to the Language and Speech Laboratory at the University of the Basque Country. We thank Valerie Hazan for providing the DiapixUK materials, and Julian Villegas, Michael Hobart and Gabriela Cavanagh for their help in annotating sections of the corpus.

References

Baker, R. & Hazan, V. 2011. “DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs”, Behavior Research Methods 43(3), 761–770.

Best, C. T. 1995. “A direct realist perspective on cross-language speech perception”. In W. Strange (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-language Speech Research. Timonium, MD: York Press, 167–200.

Blaauw, E. (1994). “The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech”, Speech Communication 14(4), 359–375.

Boersma, P. & Weenink, D. 2016: online. Praat: doing phonetics by computer [computer program]. version 5.3.51. Available at: http://www.praat.org/. (accessed May 2017).

Bolinger, D. 1989. Intonation and its uses: Melody in grammar and discourse. Stanford: Stanford University Press.

Cooke, M., King, S., Garnier, M. & Aubanel, V. 2014. “The listening talker: a review of human and algorithmic context-induced modifications of speech”, Computer Speech and Language 28, 543–571.

Council of Europe. 2001. Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge: Cambridge University Press.

Derwing, T.M., Munro, M.J., Thomson, R.I. & Rossiter, M.J. 2009. “The relationship between L1 fluency and L2 fluency development”, Studies in Second Language Acquisition 31(4), 533–557.

Ernestus, M., Hanique, M. & Verboom, E. 2015. “The effect of speech situation on the occurrence of reduced word pronunciation variants”, Journal of Phonetics 48, 60–75.

Flege, J. 1987. “The production of ‘new’ and ‘similar’ phones in a foreign language: Evidence for the effect of equivalence classification”, Journal of Phonetics 15, 47–65.

Flege, J. & Eefting, W. 1987. “The production and perception of English stops by Spanish speakers of English”, Journal of Phonetics 15, 67–83.

J. E. Flege, J. E. (1995). “Second-language speech learning: Theory, findings, and problems”. In W. Strange (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-language Speech Research. Timonium, MD: York Press, 233–277.

Gimson, A. C. 1964. An Introduction to the Pronunciation of English. London: Edward Arnold.

Granlund, S., Hazan, V. & Baker, R. 2012. “An acoustic–phonetic comparison of the clear speaking styles of Finnish–English late bilinguals”, Journal of Phonetics 40(3), 509–520.

Guion, S. G., Flege, J. E., Liu, S. H. & Yeni-Komshian, G. H. 2000. “Age of learning effects on the duration of sentences produced in a second language”, Applied Psycholinguistics 21(2), 205–228.

Harrington, J., Kleber, F. & Reubold, U. 2008. “Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study”, The Journal of the Acoustical Society of America 123(5), 2825–2835.

Hawkins, S. & Midgley, J. 2005. “Formant frequencies of RP monophthongs in four age groups of speakers”, Journal of the International Phonetic Association 35(2), 183–199.

Howell, P. & Kadi-Hanifi, K. 1991. “Comparison of prosodic properties between read and spontaneous speech material”, Speech Communication 10(2), 163–169.

Kang, O., Rubin, D. & Pickering, L. 2010. “Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English”, The Modern Language Journal 94(4), 554–566.

Kormos, J. & Dénes, M. 2004. “Exploring measures and perceptions of fluency in the speech of second language learners”, System 32(2), 145–164.

Kuhl, P. K. 1993. “Early linguistic experience and phonetic perception: implications for theories of developmental speech production”, Journal of Phonetics 21, 125–139.

Laan, G. P. 1997. “The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style”, Speech Communication 22(1), 43–65.

Munro, M. J. & Derwing, T. M. 1998. “The effects of speaking rate on listener evaluations of native and foreign-accented speech”, Language Learning 48 (2), 159–182.

Munro, M. J., Derwing, T. M. & Burgess, C. S. 2010. “Detection of nonnative speaker status from content-masked speech”, Speech Communication 52(7), 626–637.

Nakamura, M., Iwano, K. & Furui, S. 2008. “Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance”, Computer Speech & Language 22(2), 171–184.

Riazantseva, A. 2001. “Second language proficiency and pausing: A study of Russian speakers of English”, Studies in Second Language Acquisition 23, 497–526.

Rose, R. L. 2013. “Crosslinguistic corpus of hesitation phenomena: A corpus for investigating first and second language speech performance”. In Proc. Interspeech, Lyon, 992–996.

Scobbie, J. M., Lawson, E. & Stuart-Smith, J. 2012. “Back to front: a socially-stratified ultrasound tongue imaging study of Scottish English /u/”, Italian Journal of Linguistics 24(1), 103–148.

Trouvain, J., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Jügler, J., Laprie, Y., Mella, O., Möbius, B. & Zimmerer, F. 2016. “The IFCASL corpus of French and German non-native and native read speech”. In Proceedings 9th Language Resources and Evaluation Conference (LREC), Portoroz, 1333–1338.

J. Vaissière, J. 2008. “Perception of intonation”. In D. B. Pisoni & R. E. Remez (Eds.), The Handbook of Speech Perception. Blackwell, 236–263.

Van Engen, K. J., Baese-Berk, M., Baker, A., Choi, R. E., Kim, M. & Bradlow, A. R. 2010. “The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles”, Language and Speech 53, 510–540.

Villegas, J., Cooke, M., Aubanel, V. & Piccolino-Boniforti, M. 2011. “MTRANS: a multi-channel, multi-tier speech annotation tool”. In Proc. Interspeech, 3237–3236.

Wells, J. C. 1962. A study of the formants of the pure vowels of British English. Master’s thesis, University of London.

Wells, J. C. 1982. Accents of English (volumes 1 & 2). Cambridge: Cambridge University Press.

Wester, M., García Lecumberri, M. L. & Cooke, M. 2014. “DIAPIX-FL: A symmetric corpus of conversations in first and second languages”. In Proc. Interspeech, Singapore, 509–513.

Wester, M., García Lecumberri, M. L. & Cooke, M. 2015. “/u/-fronting in English speakers’ L1 but not in their L2”. In Proc. 18th International Congress on Phonetic Sciences, Glasgow.

Figure captions

Figure 1: Example pictures from the spot-the-difference task. Left and middle: fragment of a pair of Street pictures in the DiapixUK corpus; right: DiapixFL Spanish version of one member the pair where English text has been replaced by Spanish.

Figure 2: Percentages of turns consisting of speech, non-speech, unfilled pauses and filled pauses for the English and Spanish cohorts when speaking natively and non-natively. Here, and in subsequent figures of this type, the numbers represent individual speakers, and the across-speaker means are shown in the final column (error bars depict ±1 standard errors).

Figure 3: Elongated and incomplete words, measured in counts-per-minute.

Figure 4: Normalised speech rate in words-per-minute.

Figure 5: Energy, spectral tilt and percentage of voiced frames.

Figure 6: Fundamental frequency-based acoustic parameters.

Figure 7: Median F1 and F2 frequencies for the corner vowels /a/, /i/ (bold text), /u/ (italic text).

Table captions

Table 1: Participant details.

Table 2: Statistical summary of main effects for the factors L1 and nativeness and their interaction. For significant effects, columns list the F(1,18) value with η2 in parentheses, alongside an indication of significance level (***: $p < .001$; **: $p < .01$; *: $p < .05$); for non-significant effects, actual p values are shown.

Table 3: Transcription symbols and descriptions.

Corresponding author:

Dr Maria Luisa Garcia Lecumberri

Dept. Filología Inglesa

Universidad del País Vasco

Paseo de la Universidad 5

01006 Vitoria - Spain

garcia.lecumberri@ehu.es